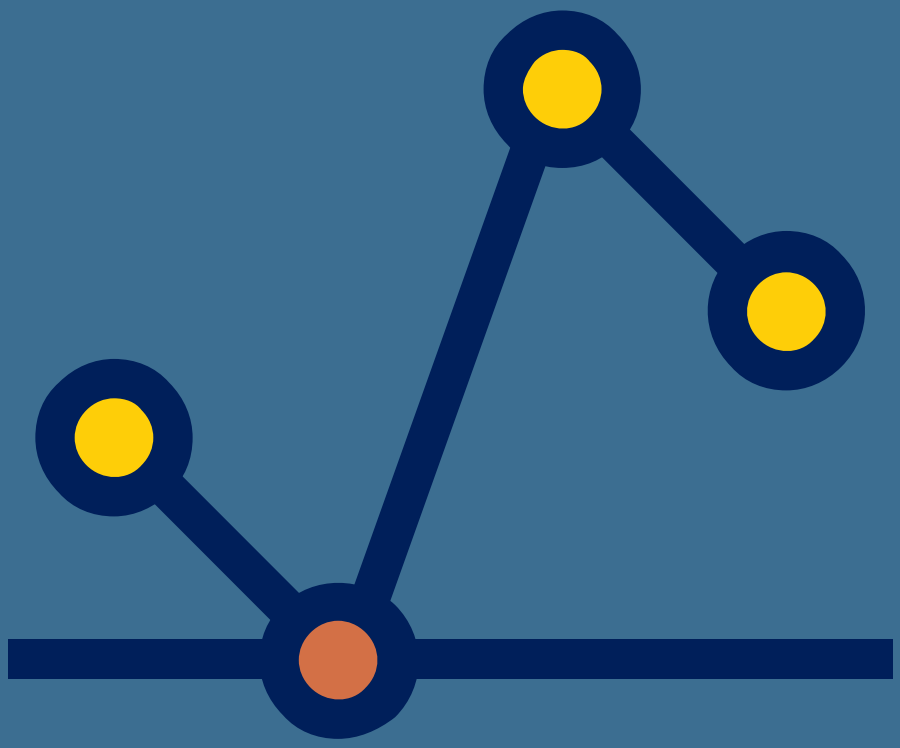
Edited by
Paola Cerchiello · Arianna Agosto
Silvia Osmetti · Alessandro Spelta

Proceedings of the Statistics and Data Science Conference



PaviaUniversityPress

Copertina: Cristina Bernasconi, Milano

Via Salasco, 5 - 20136 Milano
Tel. 02/5836.5751 - Fax 02/5836.5753
egea.edizioni@unibocconi.it - www.egeaeditore.it

Pavia University Press
info@paviauniversitypress.it – www.paviauniversitypress.it

Prima edizione: maggio 2023
ISBN volume 978-88-6952-170-6

The Role of BERT in Neural Network Sentiment Scoring for Time Series Forecast

Basili R., Croce D., Iezzi D.F., Monte R.

Abstract Sentiment scores measure the strength of customer sentiment when evaluating a product or service. This score is expressed as positive (and negative) for a numerical value between 0 and 100, where 100 is the most favourable possible result, and 0 is the least. This paper aims to combine a product's sales volume time series with the sentiment score time series of tweets generated by the BERT-NN within a state space model. We apply this model to the monthly sales volume of the Fiat L500 time series from August 2012 to Dec 2018.

Key words: BERT, ETS, Neural Network, Sentiment Scoring, state-space Model

1 Introduction

State-space models are structural models for fitting and predicting time series. Although naturally arising in multivariate contexts where the explanatory variables of a phenomenon are only partially or indirectly observable, state-space models also embody ARIMA models and can account for seasonal effect (see Durbin-Koopman

Basili Roberto
Department of Enterprise Engineering Mario Lucertini, Tor Vergata University, Rome - ITALY
e-mail: basili@info.uniroma2.it

Croce Danilo
Department of Enterprise Engineering Mario Lucertini, Tor Vergata University, Rome - ITALY
e-mail: croce@info.uniroma2.it

Iezzi Domenica Fioredistella
Department of Enterprise Engineering Mario Lucertini, Tor Vergata University, Rome - ITALY
e-mail: stella.iezzi@uniroma2.it

Monte Roberto
Department of Civil Engineering and Computer Science Engineering, Tor Vergata University, Rome, ITALY e-mail: roberto.monte@uniroma2.eu

(2012) [4] 1.2, 3.4, see also Harvey (1990) [5]). A simplified form of state-space models, the so-called ETS model, has been developed in the works of Holt (1957) [7], Winters (1960) [12], and Hyndman (2008) [8]. It is currently used in the most diverse industrial applications. The ETS model decomposes a time series into three components: a trend-cycle component, split in turn into a *level* component and a *slope* component, and a *seasonal* component. Furthermore, these components share a common innovation. The level, slope, and seasonal components are considered not observable. Therefore, the ETS model is essentially a state-space model with three non-observable state processes and the time series of interest as a single observation process.

Taking for granted the reasonably simple idea that consumer sentiment for a product influences the sales volume dynamics, the question arises of how to account for this sentiment. Referring to the ETS model, similar to Iezzi & Monte (2022) [9], we propose applying a state-space model in which we interpret one hidden component, the slope, as consumer sentiment for the product while observing a proxy of consumer sentiment, the Bert score. In contrast to the ETS approach, our goal is to use the information conveyed by the two signals to improve sales volume forecasting and, at the same time, obtain a consumer sentiment forecast. Therefore, we introduce a state-space model with three hidden state processes, one of which is consumer sentiment, and two observation processes, the time series of interest, in this paper, the Fiat 500L monthly sales volume.[1] and the Bert score proxy of consumer sentiment (see Yu et al. (2012) [13] for another approach). We use BERT in a neural network (Bert-NN) to build the proxy for consumer sentiment. Introduced by Devlin et al. in 2019 [3], BERT has rapidly become a highly regarded pre-trained neural model in the natural language processing community for its ability to tackle a wide range of language processing tasks. Its adoption by Google in 2020 further reinforced its status as a leading model in the field. BERT stands out for its bidirectional nature, which enables it to consider contextual information from both the previous and subsequent tokens in a given text. Combined with its unsupervised pre-training, BERT can effectively encode text data and generate high-quality representations that can be fine-tuned for various NLP tasks.

The remainder of the paper is organized as follows. Section 2 presents the predictive models. Section 3 discusses BERT and neural networks. In Section 4, we present the data and main results.

2 Our State Space Models

We use as a benchmark the ETS-AAA model introduced by Hyndman et al. (2008)[8]. This can be written as follows:

[1] See the website: https://www.carsitaly.net/fiat-car-sales_italy.htm

$$
\begin{aligned}
y_t &= \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t, \\
\ell_t &= \ell_{t-1} + b_{t-1} + \lambda\varepsilon_t, \\
b_t &= b_{t-1} + \beta\varepsilon_t, \\
s_t &= s_{t-m} + \gamma\varepsilon_t,
\end{aligned}
$$

where y_t is the value of the time series of interest at time t, the hidden variable ℓ_t [resp. b_t, resp. s_t] is the *level* [resp. *slope*, resp. *seasonality*] of the Holt-Winters decomposition of y_t, and the variable ε_t represents the innovation term at time t with variance σ^2. Parameter σ^2 is determined together with the parameters λ, β, γ, and the initial states of the model in the estimation procedure.
Our idea is to attribute to the slope variable b_t the role of consumer sentiment for the product (see also [9]). Of course, we cannot expect to perfectly observe such a variable, which is retained as a hidden variable, while we observe a proxy of it, which BERT-NN builds. As a consequence, in this paper, we study a state-space model where the state and observation equations take the following forms:

$$
\text{state equations } \begin{cases}
\ell_t = \beta_{\ell,\ell}\ell_{t-1} + \beta_{\ell,b}b_{t-1} + \sigma_{\ell,\ell}w_t^{(l)}, \\
b_t = \beta_{b,b}b_{t-1} + \sigma_{b.b}w_t^{(b)}, \\
s_t = s_{t-m} + \sigma_{s,s}w_t^{(s)},
\end{cases}
$$

$$
\text{observation equations } \begin{cases}
y_t = \beta_{y,\ell}\ell_t + \beta_{y,b}b_t + \beta_{y,s}s_t + \sigma_{y,y}w_t^{(y)}, \\
z_t = \beta_{z,b}b_t + \sigma_{z,z}w_t^{(z)}.
\end{cases}
$$

Here, y_t is still the observed sales volume time series of the product, and z_t represents the observed BERT-NN score on the hidden sentiment variable b_t. To add more flexibility to our model compared to the ETS-AAA model, we introduce the additional parameters $\beta_{\ell,\ell}, \ldots, \beta_{z,b}$, Moreover, we introduce independent innovations $w_t^{(l)}, \ldots, w_t^{(z)}$, with variances $\sigma^2_{\ell,\ell} \ldots, \sigma^2_{z,z}$, respectively. All these parameters, together with the initial states of the models, are estimated in a recursive procedure using the functions of the *MARSS* [6] *R* library.

3 Transformed-based Sentiment Analysis

To capture the sentiment expressed in text for use in a state-space model, we treated the semantic processing problem as a classification task and utilized the BERT neural classifier to address it. This approach is based on the work of Devlin at al. (2019) [3], and Vaswani et al. (2017) [11]. BERT provides a sentence encoding model capable of producing contextualized lexical embeddings for individual words and an encoding vector for the whole sentence. This is achieved through a *pre-training* stage applied to millions of unlabeled texts, primarily based on acquiring an expressive and robust language and text model. By stacking a dedicated network, BERT can easily be

adapted to various and diverse tasks through *fine-tuning*. Usually, a shallow multi-layer perceptron is represented by a dense layer to optimize task-specific parameters.

In this work, we adopted a fine-tuning process for BERT dedicated to sentence classification, i.e., operating on a single text given as input. It enables customization of the final classifier to suit the specific problem and fine-tun all network parameters, including those of BERT, over just a few epochs. This prevents "catastrophic forgetting" of the linguistic information gained during pre-training. It is important to note that pre-training imposes no bias on the target language. The language model learned by BERT can be acquired regardless of the language used in the input texts used in the pre-training stage. This has led to the creation of multilingual language models, such as XLM-RoBERTa, as demonstrated in Conneau et al. (2020) [2]. Here, we fine-tuned such a multilingual architecture on the SENTIPOLC dataset in Basile et al. (2021) [1] that includes tweets in Italian annotated about subjectivity with polarity labels reflecting the writer's sentiment. In the tweet analysis, we generalize the simplified view that a tweet can be either positive or negative. Instead, we acknowledge that a text can express both positive and negative polarity. Hence, two classifiers are necessary to estimate the two independent probabilities $p(\texttt{pos}|t)$ and $p(\texttt{neg}|t)$ of a message expressing a positive and negative sentiment, respectively. This allows mapping the classification task into two binary classification tasks[2]. To fine-tune the targeted sentiment analysis task, we adopt the XLM-RoBERTa architecture to generate the final hidden vector $\vec{t} \in \mathbb{R}^h$ (with h the dimensionality of the embedding space) corresponding to the first input token ([CLS]) as the aggregate representation of each micro-post. Two different classifiers are applied, one for the pos class and one for the neg one. For each class, we stacked a classification layer with weights $\mathtt{W}_{pos} \in \mathbb{R}^h$ and $\mathtt{W}_{neg} \in \mathbb{R}^h$. The output probability $p(\texttt{pos}|t)$ is estimated by evaluating the sigmoid function, i.e., $sigmoid(\vec{t}\,\mathtt{W}^{\intercal}_{pos})$ and $sigmoid(\vec{t}\,\mathtt{W}^{\intercal}_{neg})$. The two binary cross-entropy losses $\mathcal{L}_{pos}$ and $\mathcal{L}_{neg}$ are evaluated against the annotated data. The final loss is combined in a multi-task fashion as in Liu (2019) [10], i.e., $\mathcal{L} = \mathcal{L}_{pos} + \mathcal{L}_{neg}$. After fine-tuning the architecture, each tweet is processed through a BERT-based neural network, which assigns both a probability that the text is positive and a probability that it is negative[3]. Time forecasting needs to account for subjectivity and polarity signals over time. A specific aggregation method was employed to translate the classification evidence from individual tweets into signals that reflect collective information for a given period. While other methods exist, the chosen strategy is simple and easily implementable. First, probability values are used to map tweets to classes through thresholds. All tweets having $p(\texttt{pos}|t) \geq \tau_{\texttt{pos}}$ are considered positive, while all tweets $p(\texttt{neg}|t) \geq \tau_{\texttt{neg}}$. To summarize the expression of sentiment in messages $t \in T^m$ that refer to a specific time (i.e., the month m), we define the probability $p(\texttt{pos}|T^m) \approx \frac{|\{t \in T^m | p(\texttt{pos}|t) \geq \tau_{\texttt{pos}}\}}{|T^m|}$. This corresponds to the

[2] A tweet is judged fully positive when associated with $p(\texttt{pos}|t) = 1$ and $p(\texttt{neg}|t) = 0$. A perfectly negative tweet is mapped to to $p(\texttt{pos}|t) = 0$ and $p(\texttt{neg}|t) = 1$. As a result, a neutral tweet corresponds to $p(\texttt{pos}|t) = 0$ and $p(\texttt{neg}|t) = 0$, while a contrastive tweet expresses both polarities with $p(\texttt{pos}|t) = 1$ and $p(\texttt{neg}|t) = 1$.

[3] For example, t = "*Fiat 500L, la "macchina" che fa anche il caffe' @anonymizedauthor!*" is considered by the classifier to be $p(\texttt{pos}|t) = 0.72$ and $p(\texttt{neg}|t) = 0.09$.

percentage of tweets observed during m classified as positive. The same holds for the negative polarity, i.e., $p(\texttt{neg}|T^m) \approx \frac{|\{t \in T^m | p(\texttt{neg}|t) \geq \tau_{\texttt{neg}}\}}{|T^m|}$. The initial empirical evidence suggests a correlation between the target distribution and the negative polarity expressed in the messages. As a consequence, we fed the state space model with the distribution of $p(\texttt{neg}|T^m)$ with a threshold value of $\tau_{neg} = 0.7$.

4 Main Results

To test our model, we consider the FIAT 500L monthly sales volume time series, from August 1, 2012, to December 31, 2018. We collected a corpus of 20,137 tweets for the same period and measured sentiment using BERT-NN. As benchmarks, we apply the ETS-AAA model to both the sales volume and the sales volume logarithm time series. The necessity of introducing the logarithmic time series is dictated by the different scales of the sales volume time series and the sentiment signal. Table 1 summarizes the results of our analysis. Fig. 1 [resp. 2] shows the last part (from

Models	logLik	AIC	BIC	AICc	RMSE	MAE	MAPE	SMAPE %	MASE
ETS-AAA data	-606.879	1247.759	1285.738	1259.759	867.621	706.553	48.772	18.469	0.698
ETS-AAA log-data	-68.734	171.469	209.448	183.469	0.597	0.469	6.512	3.110	1.440
BERT-NN (y) log-data	7.442	9.115	35.924	11.611	0.287	0.207	2.797	1.376	0.635
BERT-NN (z) log-data diffuse initial state	7.442	9.115	35.924	11.611	0.033	0.025	35.631	19.064	0.651
BERT-NN (y) log-data	84.288	-144.575	-117.766	-142.079	0.423	0.318	4.298	2.145	0.977
BERT-NN (z) log-data random initial state	84.288	-144.575	-117.766	-142.079	0.037	0.028	35.922	21.680	0.718

Table 1 Validation measures - ETS - AAA and BERT-NN for Fiat 500L

Dec 2016 to December 2018) of the fitted and predicted results from log-scaling the ETS-AAA model [resp. from the BERT-NN model] for FIAT 500L sales volume [resp. sales volume logarithm].

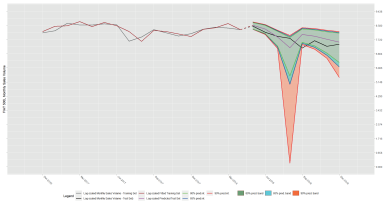


Fig. 1 Log-scaling of ETS-AAA data

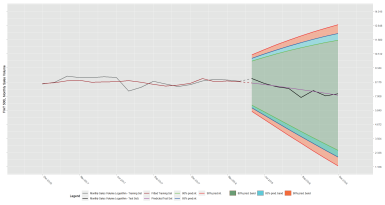


Fig. 2 BERT-NN (y) log-data

While the ETS-AAA for the original time series outperforms the ETS-AAA for the logarithmic time series, the BERT-NN with a diffuse initial state seems to

outperform both models. However, the training set fitting of the latter model presents non-optimal residuals in terms of serial correlation because it seems unable to filter the seasonal component of the time series. On the contrary, despite the higher MASE, the BERT-NN model with a randomly determined initial state performs better overall. Moreover, evaluating the performance of the BERT-NN state-space model, we should consider that we obtain not only the fit and forecast of the time series of interest but also the fit and forecast of the consumer sentiment time series. The latter might be an interesting piece of information to advise the management for a better-tailored advertising campaign.

Given the above results, we believe the state-space model approach is promising and worthy of further studies.

References

1. V. Basile, N. Novielli, D. Croce, F. Barbieri, M. Nissim, and V. Patti. Sentiment polarity classification at EVALITA: lessons learned and open challenges. *IEEE Trans. Affect. Comput.*, 12(2):466–478, 2021.
2. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*, pages 8440–8451, Online, July 2020.
3. J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
4. J. Durbin and S.J. Koopman. *Time series analysis by state space methods, 2nd ed.* Oxford University Press, Oxford; New York, 2012.
5. A.C. Harvey. *Forecasting, structural time series models, and the Kalman filter*. Cambridge Univ. Press, Cambridge u.a., 1990.
6. E.E. Holmes, J. Eric, E.J. Ward, Scheuerell M.D., and K. Wills. *MARSS: Multivariate Autoregressive State-Space Modeling*, 2021. R package version 3.11.4.
7. C.C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 2004.
8. R.J. Hyndman, A.B. Koehler, J.K. Ord, and R.D. Snyder. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Series in Statistics. Springer Berlin Heidelberg, 2008.
9. D.F. Iezzi and R. Monte. Sales forecast and electronic word of mouth: the power of feelings. In *Proceeding of the 16th International Conference on Statistical Analysis of Textual Data*, pages 489–494, Naples, Italy, July 2022. Valdistat press in coedizioni Erranti.
10. X. Liu, P. He, W. Chen, and J. Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, July 2019. Association for Computational Linguistics.
11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L . Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.
12. P.R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342, 1960.
13. X. Yu, Y. Liu, X. Huang, and A. An. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):720–734, 2012.